CHAPTER 9

WORKLOAD ASSESSMENT AND PREDICTION

Sandra G. Hart
Christopher D. Wickens

ABSTRACT

The concept of workload and its relationship to performance is introduced in this chapter. Four categories of workload measurement techniques (ratings, primary and secondary task measures, and physiological indices) are reviewed, examples of each category are described, and their strengths and weaknesses are summarized. The importance of carefully formulating the question which a measure is to address is emphasized, and it is argued that the question should guide the selection of measures. Issues relevant to implementing and interpreting workload measures are discussed and some of the reasons that different measures provide apparently conflicting information about the same situation (i.e., dissociation) are addressed. Finally, the chapter concludes with a brief description of models that can be used to predict workload.

WHAT IS WORKLOAD?

If people could accomplish all of the requirements imposed on them quickly, accurately, reliably and with little effort using available resources, the concept of workload would have minimal practical importance. However, they often cannot; in some cases task demands simply exceed operators' capabilities while in others apparently *human* limitations reflect poorly designed controls or displays, inappropriate or inadequate automation, or insufficient training. Such decrements in the performance of an individual operator, which may occur if workload is either too high or too low, can result in a reduction in overall system effectiveness. Although human adaptability and creativity are essential to the effective functioning of complex systems, human capabilities and limitations also represent a limiting factor in overall system performance. For this reason, operator workload is an important factor that must be considered in evaluating the adequacy and feasibility of operational requirements, system designs, and training procedures.

H. R. Booher (ed.), *Manprint*

Workload is a general term used to describe the cost of accomplishing task requirements for the human element of man-machine systems. This "cost" may be reflected in the depletion of attentional, cognitive, or response resources, inability to accomplish additional activities, emotional stress, fatigue, or performance decrements. Workload measures are generally obtained to evaluate the effects of different systems (or operating conditions) on any human operator or to quantify the effects of individual differences in abilities or training of specific operators working with a given system. The fact that workload varies as a function of differences between systems as well as between operators highlights the locus of the workload concept at the interface between a particular operator and a specific system. Given the number of factors that might influence workload (and which are, in turn, affected by variations in workload), specific definitions disagree about the source(s) and consequence(s) of workload and the techniques recommended for its measurement. In a single sentence, we define workload as the effort invested by the human operator into task performance; workload arises from the interaction between a particular task and the performer, as represented in Figure 9-1.

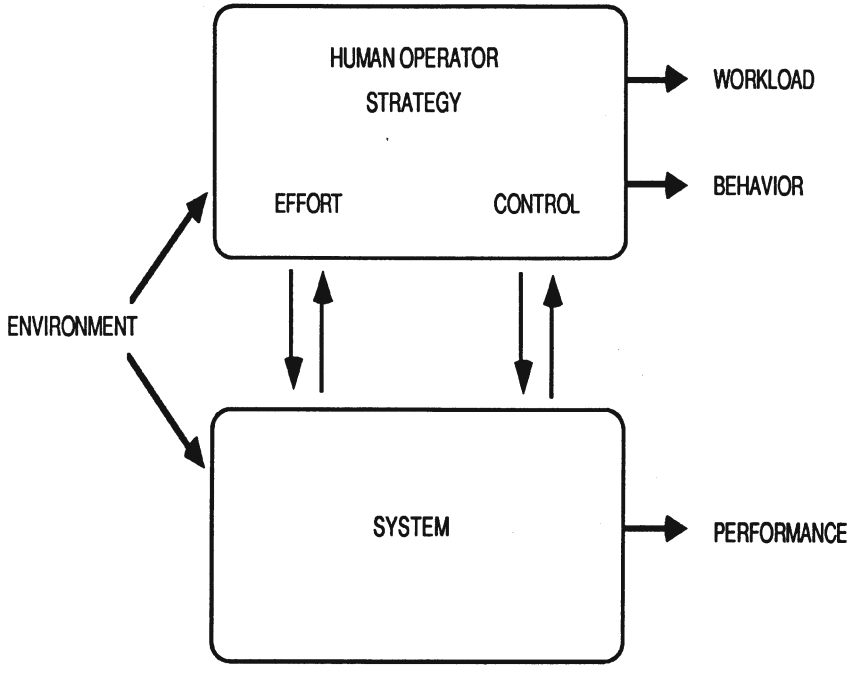

Figure 9-1
Conceptual Framework Relating Operator Performance and Workload

Although workload and performance are clearly related, the nature of this relationship is not straightforward; measures of operator and system performance and operator workload may be influenced by similar as well as different factors. In fact, operators may trade off workload and performance against each other. As task demands vary over time, operators may increase or decrease their effort (to maintain a consistent level of performance), maintain a constant level of effort (thereby allowing performance decrements to occur), defer or shed less important tasks when task demands are too high (to maintain performance on critical tasks), or complete some tasks ahead of schedule during low-workload periods (to maintain performance during high-workload periods in the future). Thus, the relationship between workload and performance from moment to moment or averaged across intervals of time depends on the strategies operators adopt and the degree to which they are able (or willing) to exert additional effort to achieve better performance, as well as on task requirements, the design of controls and displays, and environmental factors.

Formally, the relationship between workload, effort, performance, and strategies may be characterized by performance resource functions (PRF), examples of which are shown in Figure 9-2. The PRF is a hypothetical relationship that reflects the level of performance that can be obtained by a specific system, given the resources (or effort) that a particular operator invests in the performance of a specific task in a given environment (Norman & Bobrow, 1975). For example, in Figure 9-2a, performance improvements are positively, but not linearly, related to an increase in invested resources. When the PRF is steep, revealing perfect performance with few resources invested (curve B in Figure 9-2b) the system-operator interface is well designed. When it is shallow (curve A of Figure 9-2b), the system may have problems. The following hypothetical examples illustrate a number of possible relationships between workload and performance (as characterized by the PRF) to provide a framework for understanding how workload is measured and why these measurements are important.

(1) Figure 9-2b depicts the PRFs for two systems which allow operators to specify the coordinates of a point on a map. One requires a digital readout of x-y coordinates (System A). The other allows positioning of a light pen directly on the map (System B). For the light pen system (B), the operator only invests enough resources (20 percent, the vertical line) to attain maximum performance. Further investment will not improve perform-ance. Hence, lower workload and higher performance are found with System B than System A because near perfect performance can be achieved with System B by a minimal investment of resources (i.e., low effort).

(2) Figure 9-2c depicts the PRFs for helicopter flight control under day and night conditions. Equivalent performance (as measured by flight path deviations) can be achieved under both conditions when full resources are invested in flying. But, the same high level of performance can be achieved during the day with significantly less effort (as measured by an increase in

control activity) than can be achieved at night. Hence, workload is generally higher at night.

(3) Figure 9-2d depicts the PRFs obtained with two navigation displays: System A displays ownship's position and System B displays this position plus an accurate prediction of future position based on current trends. The conventional system (A) yields poorer performance, but lower workload than the enhanced system (B), the opposite relationship to that obtained in Example 1 (Figure 9-2b); the more precise information conveyed by the predictor display gives pilots more opportunity to exercise precise flight path control and encourages them to plan ahead and project the impact of their momentary control inputs into the future. Thus, they invest more resources (thereby experiencing higher workload) to exercise that control (thereby producing better performance).

(4) Figure 9-2e depicts PRFs for digital data entry obtained with either voice recognition or keyboard devices. Due to delays associated with the computer algorithm for voice recognition, performance with this system is poorer than with the keyboard. But, the use of voice, a more natural output channel than keyboard entry, requires the investment of fewer resources. Hence, the effort associated with using this system is lower.

(5) Figure 9-2f depicts PRFs that represent the improvement in pilots' skills that occur during simulator training. Such data might be used to determine when a pilot should transition to the actual aircraft. On day 4 (curve A) the pilot's routine performance is good, but he would be incapable of handling an in-flight emergency without sacrificing flight control. On day 6 (curve B) flight performance has not improved further, but the pilot now has adequate resources available to handle the emergency. As in Figure 9-2b, flight performance is equivalent, although the degree of effort that must be invested to achieve that level of performance is different.

(6) Figure 9-2g depicts the performance of two potential air traffic controllers. After extensive training, Operator A has more total resources available to allocate to the task, as predicted from basic tests of intelligence (Wickens & Weingartner, 1985). Thus, Operator A will be able to outperform Operator B in both single and dual task performance when full resources are invested, as well as when some resources are diverted to a concurrent task.

The previous examples all represent situations actually encountered in the field of system design, operator training, and assessment. They illustrate the interrelationships among the concepts of workload, performance, and effort and suggest why system performance is not a concept that can be considered in isolation from operator workload. In the following section, how workload can be assessed in operational environments and the role of operator strategies in that assessment will be considered. Then, some methodological problems related to dissociations among measures will be considered. Finally, the issue of workload prediction will be addressed.

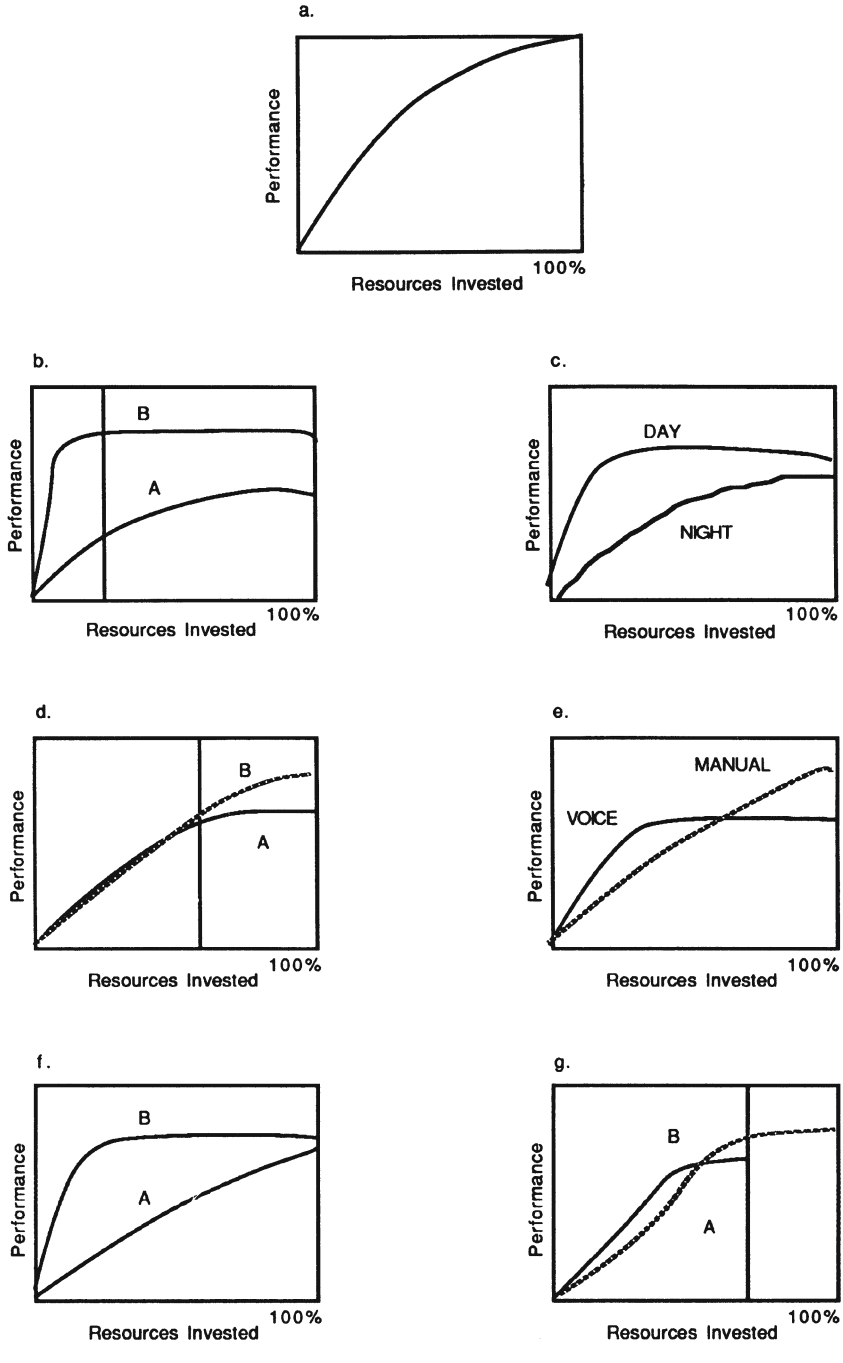

Figure 9-2
Different Examples of the Performance-Resource Function

HOW IS WORKLOAD ASSESSED?

Workload is clearly a complex phenomenon – a constantly changing, multidimensional target for analysis and assessment. The range of questions that might be asked and practical constraints further add to the complexity. However, a number of valid and practical measures have been developed that can quantify different aspects of the workload experienced by operators performing a wide variety of activities in very different environments. The measures, which will be described below, are distinguished by their objectivity, sensitivity, diagnosticity, and practicality. Since the type and quality of information provided by available measures differs, it is always a good practice to use multiple measures to develop a complete workload profile and to obtain converging evidence from different sources.

Numerous documents exist that provide detailed descriptions of available measures, the situations in which they have proven to be useful, and information about how to implement them. A limited list might include: Gopher and Donchin (1986), Hancock and Meshkati (1988), Hart (1986), Lysaght et al. (1989), Moray (1979), Moray (1988), O'Donnell and Eggemeier (1986), Roscoe (1987), Stassen, Johannsen, and Moray (1988).

Techniques used to assess operator workload generally fall into four categories: (1) ratings provided by operators or observers, (2) measures of primary task performance, (3) measures of performance on additional, "secondary" tasks introduced for the purpose of measuring residual attention or capacity, and (4) measures of covert behavior (e.g., changes in heart rate, eye blinks, eye movements, or electrical activity of the brain) which are generally referred to as "physiological" measures. Table 9-1 summarizes the categories and types of measures reviewed below.

Rating Scales

Ratings provided by task performers, observers, or experts are the most widely used measures of workload, and they often serve as the criteria against which other measures are judged. They provide an integrated summary of workload, from the perspective of the operator or an observer, and are the most direct method for evaluating the human cost of task performance. However, formally quantifying workload experiences using a structured rating scale is not a natural or commonplace activity, even though workload is experienced as a natural consequence of many daily activities. Thus, ratings might be qualitatively different than informal, spontaneous evaluations; different scales address only a subset of the factors that an individual might consider relevant and may not provide an appropriate range of alternative responses.

Table 9-1
Overview of Measures Described

Category	Examples
Rating Scales	Unidimensional Ratings Pilot Objective/Subjective Workload Assessment Technique (POSWAT) Hierarchical Ratings Cooper-Harper Handling Qualities Rating (HQR) Modified Cooper Harper (MCH) Bedford Scale Multidimensional Ratings Subjective Workload Assessment (SWAT) NASA-Task Load Index (NASA-TLX)
Primary Task Measures	Reaction Time Capture Time Tracking Error Control Variability Errors Communications
Secondary Task Measures	Reaction Time Monitoring Time Estimation Mental Arithmetic Memory Search Tracking Embedded Tasks
Physiological Measures	Heart Rate Pupil Diameter Measures of Eye Function Event-Related Potentials (ERP)

Rating scales generally consist of an ordered sequence of response categories. Labels on the scales define the correspondence between stimuli (workload experiences) and responses (rated levels). However, there is no direct relationship between values on any workload scale and specific, measurable phenomena in the physical world; "true" zero and the upper limit are undefined and intervals may not be psychologically equal. Thus, most scales provide ordinal information and index relative differences, rather than absolute levels.

Ratings may be obtained while the task is being performed, during intervals between task segments, or upon its conclusion. The former may interfere with task performance, while the latter may result in the loss of important information, unless supplemented by memory aids. The intervals of time evaluated range from several minutes to many hours, may include relatively homogeneous or diverse activities, and represent arbitrary intervals of time or meaningful units of activities. There is an obvious trade-off between the sensitivity and precision that may be achieved by obtaining ratings frequently and the possibility of task interference and rater "burn-out." Finally, since ratings are necessarily based on memory, they reflect only a subset of the information that was available during task performance (Ericsson & Simon, 1980). However, ratings obtained immediately after a task or after a considerable delay are highly correlated and their absolute values are similar.

Since workload is not completely defined by objective task demands, the experiences of different individuals faced with identical task requirements may be quite different. In addition, people may consider different variables when providing a rating (because their personal definitions of workload vary) and express different subjective biases (which may not be related directly to workload). This results in the primary drawback of subjective ratings: relatively high between-rater variability.

Although ratings are generally obtained from the person actually performing a task, observers, who can mentally project themselves into the situation experienced by the operator, can provide useful information about workload. Observer ratings may serve as the basis for operator debriefings and provide additional information that a busy operator might miss or forget to report. However, observer ratings are limited to the assessment of observable actions, task requirements, system performance, and environmental factors; it is difficult for observers, no matter how familiar they are with the task and environment, to infer the mental effort, stress, and psychological consequences of performing a task. This limits the information observer ratings can provide.

Several rating scales have been developed that have provided valuable information in a wide variety of applications. These scales may be grouped into three categories: (1) unidimensional; (2) hierarchical (Modified Cooper Harper [MCH], Bedford); and (3) multidimensional (National Aeronautics and Space Administration-Task Load Index [NASA-TLX], Subjective Workload Assessment Technique [SWAT]).

Unidimensional Ratings

Global ratings are easy to obtain (a single number can be given verbally during performance of almost any task) and provide a convenient summary value. However, unidimensional ratings provide no diagnostic information

and between-rater variability is generally high (Hart & Staveland, 1988; Byers, Bittner, & Hill, 1989) – different raters base their evaluations on different aspects of the situation. There is no standard format for unidimensional ratings, although most require raters to provide numeric values (using scales that range from 1-7, 1-10, or 1-100), descriptive labels (e.g., very low, low, moderate, etc.), or magnitude estimates marked on scales presented on paper or a computer screen that are later converted to numeric values. A portable system, the Pilot Objective/Subjective Workload Assessment Technique (POSWAT) has been developed with which global workload ratings can be obtained and recorded in-flight (Mallery & Maresh, 1987) using an array of 10 labeled buttons.

Hierarchical Ratings

The Cooper-Harper Handling Qualities Rating (HQR) Scale (Cooper & Harper, 1969) was one of the earliest rating scales developed and is still used widely. Although it was developed to obtain subjective evaluations of aircraft handling qualities, subsequent applications have shown that it is also sensitive to many of the same factors that also influence workload. Raters make a series of decisions, each of which discriminates between two or three alternatives. Each decision leads the rater to another choice or to a final numeric rating ranging from 1-10. Raters may read the scale each time they provide a verbal or written rating or do so from memory. Several modified versions of the scale have been developed that retain the decision tree format and 10-point scale, but substitute terms that address workload more directly.

The Modified Cooper-Harper (MCH) Scale (Wierwille, Casali, Connor, & Rahimi, 1986) has been tested in the laboratory, simulated flight, and Army field tests of ground-based systems. Although it has been found to be sensitive to workload variations in simulated flight, it has proven to be less useful in other environments.

The Bedford Scale (Roscoe & Ellis, in press) was explicitly developed for use in flight. The wording of the scale is focused on spare capacity. It has been used widely in England and Europe to evaluate pilot workload in military and civilian aircraft and in simulation and flight research in the United States, but has received limited use in non-aviation environments.

The advantage of these and other decision tree formats is that they separate the evaluation process into a series of explicit decisions. These scales are easily implemented and scored and can be used without creating unacceptable interference in even demanding operational environments. However, they do not provide diagnostic information and have not received the extensive evaluation and application that other scales have received.

Multidimensional Ratings

Two of the most commonly used rating scales, SWAT and NASA-TLX, involve a procedure whereby ratings on several subscales are combined to derive a summary score. These scales are based on the assumption that people can evaluate component factors more reliably than they can the global concept. One of their greatest strengths is that they provide diagnostic information about the specific source of a workload problem, as well as a global summary.

SWAT (Reid, 1985; Reid & Colle, 1988; Reid & Nygren, 1988) consists of three subscales: time load, mental effort load, and psychological stress load. Each dimension is represented by a three-point rating scale (1=low, 2=medium, and 3=high). The 27 possible combinations of three levels of each of three scales are presented on individual cards for subjects to rank from lowest to highest workload prior to an experiment. These rankings are used to create a 100-point, unidimensional scale that has interval properties. Each combination of ratings on the three subscales has a unique position on this scale which is assigned a numerical global workload value. The sorting process, although time-consuming, is valuable as it provides an interval scale and considers individual differences in workload definition. The limited number of dimensions and interval scale values make SWAT attractive for use in operational settings; however, this limited range of allowable ratings (e.g., low, medium, and high) also reduces the sensitivity of the scale. The primary drawbacks are relatively high between-rater variability and low sensitivity in situations where overall workload levels are generally low (see, for example, Reid, 1985). SWAT has been used successfully to evaluate workload in the laboratory, simulation and flight research, and various ground-based operations. Thus, there is accumulated evidence that SWAT provides valid results and can be implemented in most environments.

NASA-TLX provides an estimate of overall workload based upon a weighted average of six subscale ratings: mental demand, physical demand, temporal demand, own performance, effort, and frustration (Hart & Staveland, 1988). Subscale ratings, which range from 1-100 in 5-point increments, are given verbally or by selecting a position along a scale presented on a rating form or computer screen. In addition, raters quantify the relative importance of each factor in creating the workload they experienced. These values, which range from 0 to 5, are used to weight the magnitude ratings when computing the overall workload score. Diagnostic information is provided by variations in subscale ratings as well as the weight given to each factor. NASA-TLX has been used successfully in environments ranging from the laboratory to military and civilian helicopter, general aviation, transport, and military jet simulators and aircraft, and in ground-based systems. Between-rater variability is consistently lower than typical of other rating scales, significant correlations with other measures of workload and performance are generally found, and subtle as well as gross workload differences have been discriminated throughout the workload range.

Summary: Rating Scales

A 10-point unidimensional scale, SWAT, NASA-TLX, and MCH were compared in a series of field tests conducted for two Army air defense systems: (1) the Line of Sight Forward (Heavy) (LOS-F-H) Forward Area Air Defense System (FAADS) (Hill, Zaklad, Bittner, Byers, & Christ, 1988) and the Pedestal Mounted Stinger (Byers, 1989) and for the Aquila Remotely Piloted Vehicle (Byers, Bittner, Hill, Zaklad, & Christ, 1988). Ratings were obtained, for at least some of the scales, either: (1) prospectively – raters familiar with the basic system were asked to evaluate the *potential* impact of a hardware modification, different tactical situations, or a reorganization of crew functioning (Hill, Byers, Zaklad, Bittner, & Christ, 1989); (2) during or soon after the field tests; or (3) retrospectively – many months after initial field experience, operators and subject matter experts rated the "generic" workload associated with mission conditions, task segments, and operator position (Bittner, Byers, Hill, Zaklad, & Christ, 1989). Generally, high correlations were found among the four scales, suggesting that they all reflect the same "overall workload factor." Many of the scales were able to distinguish meaningful workload differences among mission segments, variations in task demands, and crew position. However, NASA-TLX ratings were more closely related to objective measures of performance (good performance was generally associated with low workload), they had the lowest between-rater variability and the highest overall workload factor validity across tests, and the best user acceptance (see, for example, Byers et al., 1988). Although NASA-TLX was considered to provide the most accurate description of workload and was recommended for use in field evaluations of Army systems, the unidimensional ratings, which consistently out-performed SWAT and MCH, were also recommended for use in screening gross levels of workload, in preparation for a more diagnostic evaluation of workload problem areas (Hill et al., 1988). Although the unidimensional scale did not provide the diagnostic information given by the NASA-TLX subscales, it took considerably less time and effort to use than did NASA-TLX.

Rating scales are the most practical and generally applicable measure of workload. They are easy to implement and score, appropriate in almost any environment, acceptable to most operators, and have a certain amount of face validity. The rank ordering of ratings is quite stable across raters, although the absolute values of the ratings exhibit relatively high variability. In addition, subscale ratings can provide diagnostic information. However, ratings can represent no more that the rater's memory of what was experienced, integrated across time. Thus, most ratings are insensitive to momentary variations in workload and they are subject to rater biases. Although improvement in the psychometric properties of these scales is warranted, their practical utility and the wealth of information they provide outweigh their drawbacks.

Primary Task Measures

Although improving performance may not be the only motivation for measuring workload, measures of performance are an essential component of any workload analysis; it is difficult to interpret workload measures without knowing the level of performance the operator was able to achieve. In addition, some performance measures can provide objective answers to workload questions. Since ensuring adequate performance is the motivation behind most applied workload analyses, it seems reasonable to consider performance measures first when selecting a workload metric. However, as many of the examples in Figure 9-2 illustrate, these measures reflect what the man-machine system was able to accomplish, rather than the cost of doing so for the human operators, and they are also influenced by factors other than workload (e.g., system response time). In addition, accurate measures of performance may not be available in field evaluations. For highly automated systems, system performance depends on the operator's inputs at a very high level; moment to moment variations depend entirely on the performance of the automated system, while operator inputs occur infrequently.

Types of Primary Task Measures

There are three classes of performance measures: (1) Accuracy (number of correct responses, control error compared to a target value), (2) Speed (response time measured in seconds or fractions of seconds), and (3) Number (how many tasks or task elements were completed correctly within an interval of time). Some measures summarize the effectiveness of the operator's activities, while others also provide information about the fine-grained structure of the operator's control strategies. The former reflect the combined output of the operator's behavior and system output, while the latter measure operator effort (and, thus, workload) more directly.

Limitations of Primary Task Measures

For the purpose of workload assessment, the usual assumption is that decrements in performance indicate higher workload; more errors, longer response times, higher control error, and fewer tasks completed are taken as evidence of increased workload. However, the actual relationships between workload and performance are more complex. O'Donnell and Eggemeier (1986) suggested that: (1) for relatively easy tasks, consistent performance is maintained over a range of difficulty levels (although workload increases), (2) for moderately difficult tasks, performance deteriorates linearly as task demands increase (and workload increases), and

(3) for very difficult tasks, operators may ignore some tasks and maintain a consistent level of effort on others, even in the face of increasing demands, so that performance deteriorates, but workload does not increase further. In general, measures of performance appear to correlate with task demands and workload for moderately difficult tasks only. For very easy or very difficult tasks, measures of workload and performance dissociate.

Numerous models of attention and performance have been proposed to explain the relationships found between task demands, workload, and performance (e.g., O'Donnell & Eggemeier, 1986; Vidulich, 1988; Wickens & Yeh, 1988), as suggested by the examples in the beginning of the chapter. The basic assumption is that the resources required to perform tasks are available in finite amounts. As difficulty is increased, more resources are required to maintain consistent performance. If available resources are sufficient, performance will remain constant. If they are not, performance will degrade on a single task or on one or more concurrent tasks. Since some tasks have been found to interfere with each other more than others, Wickens (1980) and Wickens and Liu (1988) suggested that the performance of concurrent tasks depends on the pattern of requirements that the two tasks share in common. Thus, performance decrements on one task may occur in the presence of some concurrent tasks, but not others, due to the pattern of resource competition between the two tasks. However, it is difficult to assess workload from the measures of performance that are obtained. If two tasks do not compete for the same resources, or if sufficient resources are available to perform both, performance is maintained but workload is increased. If two tasks do compete for the same, insufficient resources, performance on one or both tasks will degrade, but workload may not be affected.

Performance is also influenced by task schedule; almost any task can be performed well if unlimited time is available and even easy tasks may become impossible to perform if the time available is reduced below a critical point. Thus, measures of performance will reflect workload only in the region where sufficient, but not unlimited, time is available for task performance; as time pressure is increased within this region, measures of performance will show a decrement even though additional effort is exerted.

In both theoretical and applied research, it is always assumed that operators attempt to respond immediately and perfectly to task demands. If they do not, then these measures lose their meaning. Even for simple tasks, an operator may choose to emphasize speed at the expense of accuracy, or vice versa. This may result in two competing estimates of his performance.

In most operational tasks, there is considerable flexibility in when and how task requirements are accomplished. Operators may try to maintain acceptable performance on each of two concurrent tasks (by time sharing or rapidly switching between them), emphasize one task at the expense of the other, or perform the tasks sequentially. In fact, delaying the performance of

one (less important) task in favor of another (more important) task might well represent an optimal strategy. In addition, most tasks do not have be performed perfectly. Rather, a vehicle must be kept within certain boundaries (not perfectly lined up) and many discrete tasks can be performed at the operator's discretion, as long as they are completed by a deadline. Thus, operators' task performance strategies determine when and how they focus their attention on individual tasks. This, in turn, determines which measures of performance are the most appropriate for analysis and the magnitudes of performance decrements that will be recorded for specific task components. Strategies adopted to minimize workload might result in poor performance on one element of a complex task, but, if the strategies are part of a more global workload-management strategy, they can result in better overall performance. Alternatively, very good performance, often thought to index low workload, might also reflect the exertion of extreme effort and very high workload. "Extreme effort" might be defined as a level of exertion that can be sustained for only a short interval of time.

Primary Task Measures for Simple Tasks

Despite these problems, several types of performance measures seem to covary with other measures of workload sufficiently reliably that they might be considered. For example, in laboratory research, consistent relationships have been found between an increase in reaction time (due, for example, to an increase in the number of remembered items, alternative choices, or arithmetic operations) and other workload measures (e.g., Mosier & Hart, 1985). For target acquisition tasks, a reliable relationship has been found between increased capture time (as target size decreases or its distance increases) and subjective ratings (e.g., Hart, Shively, Vidulich, & Miller, 1986). For tracking tasks, including simulated flight, increased tracking error (associated with an increase in the difficulty of the forcing function, the number of controlled axes, or the order of control) generally correlates with an increase in subjective ratings and decrements in secondary task performance (e.g., Bortolussi, Hart, & Shively, 1989; Kramer, Sirevaag, & Braune, 1987; Vidulich & Wickens, 1986).

Primary Task Measures for Complex Tasks

Interpretation of performance measures is quite complicated when many measures are available; different measures taken concurrently might suggest very different levels of workload. Since the units of measurement, frequency of occurrence, and priority vary across task components, it is not always easy to combine multiple performance measures into a summary

figure of merit for performance which might, in fact, reflect workload. Two approaches to such combinations are possible: (1) The use of multivariate analyses (i.e., principle components analysis) that reveal the most important variables in discriminating among levels of workload, and (2) A linear combination of all of the measures of primary task performance that the investigator feels are important, such that measures of "good" or "bad" performance, respectively, are consistently given the same sign. Although, in theory, all measures that are recordable could be combined, there are some important qualifications to using some measures. For example: (1) Errors generally provide a poor index of workload; slips and blunders occur as often in low workload situations as high (Morris & Rouse, 1988). However, the activities required to resolve an error can contribute to a subsequent increase in workload; (2) Operators must devote at least some attention to a task for its performance to reflect overall workload; (3) The measure must be recorded with adequate frequency and accuracy; and (4) The operators' inputs must not be masked or delayed by the system. Figures 9-2d and 9-2e depict explicit examples of situations where higher workload is associated with better, rather than worse primary task performance in a complex system.

Some elements of complex primary tasks do appear to provide consistent and reliable indices of workload. For example, operators' abilities to estimate the passage of time generally degrades as overall workload (as indexed by other measures of workload) is increased in laboratory, simulation, and flight research. Altitude control variability often increases in simulation and flight experiments as other task demands are increased. Communications are often delayed and abbreviated as overall task demands increase.

Summary: Primary Task Measures

While measures of performance on the task of interest can be used to estimate momentary variations in workload and the degree to which a specific level of effort can be sustained, the problems outlined above suggest caution in their use. In fact, Gopher and Donchin (1986) concluded that direct measures of task performance are usually poor indicators of mental workload because they do not reflect the resource investments prompted by changes in task demands and do not diagnose the source of load. Furthermore, a review of the field recently conducted for the Army Research Institute (Lysaght et al., 1989) concluded that primary task measures "should not be generally treated as appropriate for assessments of overall workload." Taken together, these recommendations provide a strong note of caution about using primary task measures to evaluate workload, although they must always be obtained to determine whether or not the operator was able to accomplish the task; other measures of workload are virtually impossible to interpret without such information. Thus,

while primary task performance measures can evaluate the "bottom line" – Can the task be done? – they may not reflect the workload that an operator experienced in achieving that level of performance. As task demands are increased, operators are often able maintain a consistent level of performance (which would suggest no difference in workload), even though the workload "cost" of doing so may be greater (if they responded to the demands of the task by exerting additional effort) or less (if they responded by adopting a more efficient task-performance strategy). Conversely, performance decrements accompanying an increase in task demands may either reflect a constant or reduced level of effort (i.e., the operator did not choose to devote more resources to the task or adopt a more efficient strategy) or the upper limit of his capabilities (i.e., the operator did not have any additional resources available and a more efficient strategy was not possible).

Secondary Task Measures

Because measures of performance on the task of interest may not provide an adequate estimate of workload (if task demands remain within the operator's capabilities, more effort may be exerted to maintain consistent performance), yet objective measures are desirable, the use of "secondary" tasks was developed as an alternative approach. Operators are instructed to maintain performance on the primary task and use their "reserve capacity" to perform the additional tasks. The secondary tasks impose a sufficient additional load so as to exceed the operator's capabilities. The level of performance on these tasks is used as an indirect measure of the resources demanded by the primary task; secondary task performance degrades as primary task demands increase. Thus, secondary tasks can provide a useful measure at the low end of the workload continuum, where primary task performance measures are insensitive.

Attractive as the concept seems to be, in theory, secondary tasks are not without some problems. For example, if primary task demands are fairly high, particularly in an operational environment, the operator may simply abandon the secondary task in order to maintain acceptable performance on the primary task, making the secondary task measure ineffective. Initially, it was thought that secondary task "yardsticks" could be developed and used to compare the workload of a variety of primary tasks (Ogden, Levine, & Eisner, 1979). However, the data suggest that secondary tasks are selectively sensitive, depending on the pattern of requirements they share with the primary task, again suggesting the existence of multiple resources (Wickens, 1980, 1984). Each task requires a specific pattern of resources. Concurrent tasks that require similar (insufficient) resources will interfere with each other, while those that require different resources will not. Because interference results in performance decrements on one or both tasks, the assumption that "secondary" tasks are always performed after "primary" task

requirements have been met is not always supported. Another possible problem is that most secondary tasks occur at discrete intervals. Thus, their performance reflects workload at the times they were introduced, rather than the workload throughout a period of time. If secondary task presentations are time-locked to significant events in the primary task, however, they can provide more precise information than can a more global measure (Kantowitz, Bortolussi, & Hart, 1987). If operators modify their primary-task performance strategy when a secondary task is present, the information that a secondary task provides about primary task workload becomes ambiguous.

A careful task analysis is particularly important in using secondary tasks; tasks must be selected that require the same resources necessary for performance of the primary task. Since every task requires several types of resources, performance on several tasks that require different patterns of resources can provide converging information about primary task workload. In addition, using secondary tasks that have graded levels of difficulty can provide a more accurate estimate of primary task workload.

Most secondary tasks are relatively simple activities for which the input (visual, auditory) and output (verbal, manual) can be presented precisely and measured directly and accurately. The intervening cognitive processes are predicted by psychological models and inferred from variations in the speed and accuracy of performance. Most tasks were originally designed for purposes other than workload assessment (e.g., to test theories of human performance, memory, and attention), however, others represent simplified versions of "real-world" tasks. The following describes several of the secondary tasks which have received the widest application:

Reaction Time

Reaction time tasks generally include a visual or auditory stimulus presented during performance of a primary task. The operator responds by pressing a button or making a verbal response. Multiple-choice tasks are more sensitive than single-choice tasks, as they impose some information processing and response selection load. Increased response time and errors index an increase in primary task workload (e.g., Bortolussi et al., 1989; Kantowitz et al., 1987). Depending on the modality of input (visual/auditory) and output (verbal/manual) selected, this task can be designed so as to require the same resources also required by a specific primary task.

Monitoring

Monitoring tasks generally require the operator to pay consistent attention to one sensory modality (visual/auditory) and respond when a particular

event occurs (verbally/manually) or maintain a running count (which adds a memory requirement). Increased response time, misses, and false alarms index an increase in primary task workload (e.g., Kramer, Wickens, & Donchin, 1983). Again, this task can be implemented so as to compete for the resources also required by the primary task.

Time Estimation

Timing tasks require the subject to produce a specific interval of time (usually in the range of 1-20 seconds) by manually pressing a button or verbally indicating its beginning and end. As attention is drawn away from timekeeping, the length of produced durations increases (e.g., Hart, McPherson, & Loomis, 1978). Although the demands of this task are almost entirely on central processes, it has been found to be sensitive to variations in perceptual/motor load as well.

Mental Arithmetic

Mental arithmetic tasks also require primarily cognitive resources. The difficulty of the task can be manipulated by increasing the number of digits or the number of operations that must be performed (e.g., Roscoe & Kraus, 1971). Performance is measured by response latency and the accuracy of verbal, written, or typed responses.

Memory Search

Memory search tasks require the operator to remember one or more letters, number, or words (the memory set) and then to respond whether subsequent stimuli (probes) were members of the memory set or not. As memory set size increases, or as concurrent primary task demands increase, response time generally increases (e.g., Wickens, Hyman, Dellinger, Taylor, & Meador, 1986). This task imposes loading on short-term memory, as well as perceptual and response resources.

Tracking

Tracking tasks can provide a continuous index of primary task workload, although they are often impractical in operational environments. Different difficulty manipulations (e.g., forcing function amplitude and bandwidth, order of control, and number of axes controlled) create different patterns of interference with concurrent primary tasks (e.g., Jex & Clement, 1979). In

general, error about a target value increases as the difficulty of a concurrent task is increased.

Embedded Tasks

One problem with secondary tasks in operational environments is that they tend to be considered unimportant and uninteresting by the operators and their performance may be terminated altogether. A solution to this problem is to present embedded secondary tasks which are designed to appear as a natural and integral part of the task of interest. This method of presenting a secondary task substantially improves user acceptance (in operational situations) and increases the likelihood that operators will attempt to perform them. Many of the tasks described above can be modified so as to integrate them with an operational task. For example, a memory search task can be designed as a response to a radio call sign. The time at which a particular activity is performed can be treated as a time production. Discrete responses to alternative display configurations can be evaluated as reaction time or monitoring tasks, and so on. In addition, there are many components of complex tasks whose performance reflects overall workload levels, such as communications, which can be singled out for analysis.

Summary: Secondary Task Measures

Performance on a concurrent, additional task can index the reserve capacity remaining after primary task performance and provide diagnostic information about the specific resources required. However, secondary task performance may interfere with or change primary task performance. Thus, this measure may be inappropriate for many operational situations unless presented as embedded elements of the primary task. In addition, not all secondary tasks are equally sensitive to the specific types of demands imposed by different activities. A lack of secondary task decrement can never be interpreted with certainty as evidence of low workload; it might just indicate that the resources required by that secondary task were not depleted by the performance of that primary task. Thus, the use of several secondary tasks, each of which demands different combinations of resources, will provide a more accurate assessment of the sources and magnitudes of primary task workload. A thorough task analysis and a strong theoretical basis is necessary to select and implement secondary tasks successfully.

Physiological Measures

Measurable, involuntary changes in such measures as heart rate, eye blink rate, pupil diameter, respiration, blood pressure, electrical activity of the

brain, and so on may accompany variations in the physical and mental demands of a task (Hancock, Meshkati, & Robertson, 1985; Wierwille et al., 1986). In addition, since emotional stress may accompany an increase in workload, measures of arousal provide an indirect indicator of workload. Thus, many physiological measures have the potential for providing an objective and unobtrusive indication of operators' responses to the demands placed on them. In addition, these measures rarely require an overt response, and, thus, do not interfere with task performance. Finally, many physiological indicators can be monitored and measured continuously, thereby providing a fine-grained analysis of subtle, momentary changes in workload.

There are two classes of physiological measures: (1) measures of emotional and physical activation (e.g., heart rate, pupil size), and (2) measures of mental and perceptual processing (e.g., event related potentials and direction of gaze). Measures of voluntary actions (such as hand or leg movement) are not generally considered to be "physiological" measures.

Although the selection of a particular physiological measure to index workload is rarely based on sound theoretical considerations, most of the measures have a sound basis in physiology. The general expectation is that heart rate will increase, heart rate variability will decrease, pupil diameter will increase, the pattern of eye blinks and eye movements will change, and the amplitude of event-related potentials (in response to a secondary task probe) will decrease as workload is increased. However, different measures are specifically sensitive to different types of workload. For example, heart rate is particularly responsive to stress and physical effort, while other measures of cardiovascular functioning and event-related potentials reflect mental effort. The following briefly describes a few of the physiological measures that seem most useful:

Heart Rate

Measures of the operator's heart rate provides an integrated index of the overall effect of task demands and the operator's emotional response to them. Heart rate is particularly sensitive in situations where stress and responsibility play an important role. For example, heart rate typically increases during takeoff and landing in simulation and flight research for the pilot responsible for controlling the aircraft, but does not for other pilots in the cockpit (e.g., Hart & Hauser, 1987). Heart rate is generally less useful in low-workload situations and non-operational environments. The variability of beat-to-beat intervals also indexes workload; normal heart rate irregularities are suppressed as task demands, particularly cognitive demands, are increased (e.g., Derrick, 1988; Vicente, Thornton, & Moray, 1987). Although this measure can be computed without any additional cost,

it has not proven to be as reliable a measure as is heart rate. Heart rate is easy to record in almost any environment using portable devices.

Pupil Diameter

Variations in the diameter of the operator's pupil provides an accurate index of cognitive load (e.g., Beatty, 1982). However, this measure is quite difficult to obtain reliably in operational environments, due to its sensitivity to ambient illumination. In general, the finding has been that pupil diameter increases as cognitive workload is increased.

Measures of Eye Function

Most modern systems present a substantial amount of information visually and operators may control the flow of this information consciously or unconsciously through blinking, redirecting their gaze, or changing their focus. Thus, measures of these functions can provide useful information about visual and cognitive workload. The timing, frequency, and duration of eyeblinks have been found to reflect workload-related phenomena (e.g., Stern & Skelly, 1984); blinking is inhibited until information is obtained and tends to occur after decisions are made. Information about where the operator is looking, and for how long, and transition patterns among displays provide valuable information about task-performance strategies (e.g., Harris, Glover, & Spady, 1986); operator's abilities to monitor an added visual display can indicate the visual monitoring and processing, resources required by the primary task, fixation durations may decrease, and scan patterns may change under high workload. A modification of this technique, useful in Army combat or vehicle control environments, is to calibrate workload by the reduction in peripheral scanning or head movements that result when central load increases. This may be estimated by head movements in conjunction with peripheral eye movements. The primary drawbacks of measures of eye function are the difficulty of implementing them in operational environments and interpreting the data (e.g., people do not always "see" what they are looking at and can process information from the periphery).

Event-Related Potentials (ERP)

ERPs have been proposed as an objective index of mental workload because of the obvious connection between the brain and behavior (Gopher & Donchin, 1986; Donchin, Kramer, & Wickens, 1986). The occurrence of sensory events is related to predictable patterns of electrical

activity in specific parts of the brain that can be measured by electrodes placed on the scalp. The general practice is to average the data recorded from several presentations of similar stimuli so that the ERP "signal" can be detected in the "noise" of ongoing brain activity; this is possible because the latencies and amplitudes of ERP waveforms are similar for the same types of events. Generally, cognitive workload is reflected in the amplitude of a positive component of the ERP waveform that occurs between 300-500 milliseconds after presentation of a stimulus; ERP amplitudes increase as mental effort increases and are reduced if cognitive resources are also demanded by other concurrent tasks. Although relatively difficult to implement in an operational environment, Kramer et al. (1987) found reliable relationships among event-related potential amplitudes, performance on a primary, simulated flight task, secondary task performance, and subjective ratings. Futhermore, Kramer et al. (1983) used the ERP to index workload changes as training progressed in a complex task. This measure provides a direct index of information processing activities, and, because it is time-locked to a specific event, it can provide very precise information about the workload at that time. It cannot, however, reflect continuous workload variations and is more expensive and difficult to implement and analyze than heart rate or rating measures. Finally, event-related potentials are subject to artifacts (e.g., eye movements or blinks) which are themselves related to workload and require special filtering to remove.

Summary: Physiological Measures

Some physiological measures can index the physical and emotional responses that accompany an increase in workload, while others index cognitive load. Many vary with sufficient frequency to be sensitive to momentary shifts in workload. Most can be obtained with minimal primary task interference. They provide "objective" information that is quantified in physical units. Despite these advantages, however, their use in operational environments has met with only limited success, and, with the possible exception of heart rate, no single measure covaries with other measures of workload or performance consistently enough to be recommended as the only measure of workload. As part of a battery of measures, however, they can provide useful, unobtrusive information that may be absent from subjective ratings or measures of primary or secondary task performance.

HOW ARE MEASURES SELECTED AND IMPLEMENTED?

Several expert systems have been developed to aid in selecting and applying available assessment techniques. The Workload Consultant for Field Evaluation (WC FIELDE) was developed at NASA (Casper, Shively, &

Hart, 1987) to serve as an aid in formulating the research question, assessing the research environment, evaluating the relevant task dimensions, and selecting the most appropriate measures. This system includes an extensive data base that describes the measures, evaluates their strengths and weaknesses, summarizes previous applications, and suggests how to implement them. A second system was developed by Analytics, Inc., under contract to the Army Research Institute for use during all stages of system specification, design, test and evaluation, and operation for military system acquisitions. The Operator WorkLoad KNowledge-based Expert System Tool (OWLKNEST) includes predictive techniques as well as assessment techniques in its data base (Harris, Hill, & Lysaght, 1989). It provides a brief summary and evaluation of each measure and predictive technique.

In selecting the measure(s) that will be used, the following should be considered: (1) the focus of the research question, (2) the grain of analysis that is required, (3) the probable level and sources of workload, (4) the importance of diagnostic information, and (5) practical constraints imposed by the research environment. Different measures are particularly sensitive to specific workload dimensions and magnitudes, and they may either summarize the overall effects of task performance on the operator or provide diagnostic information to isolate the causes or consequences of a specific problem. Depending on which measure is selected and how it is implemented, the data may provide a fine-grained analysis of the time-varying characteristics of the operator's behavior and experience or an integrated summary across sources of workload and time. Finally, some measures can be obtained in almost any environment with minimal cost and interference, while others are difficult to obtain in the field or require expensive equipment and special training. In practice, there is usually a trade-off between the time and effort devoted to workload assessment and the quality and precision of the information that is obtained.

Formulating the Question

The potential causes (e.g., mission requirements, equipment design, environment) and consequences (e.g., performance decrements, physical and emotional stress) of inappropriate levels of operator workload encompass a broad range of factors. Thus, a formal statement of the research question will ensure that the most appropriate measures are selected and that the obtained evidence does, in fact, answer the original question. The following represent a sample of the questions that might motivate a workload analysis: (1) Are sustained or momentary workload levels consistent with acceptable system performance? Operator's emotional and physical well being? (2) Could crewmembers perform additional duties? (3) Should crew duties be reassigned to distribute

workload more evenly? (4) Could the crew complement be reduced? (5) Will adequate performance be possible under degraded weather? Reduced visibility? Battlefield conditions? If system components fail? (6) Will the addition of automated subsystems reduce the workload of existing crewmembers? Allow a reduced crew complement? or (7) What is the best design alternative? Visual/auditory displays? Manual/vocal controls? Multifunction/integrated/separate display formats? Different measures are particularly sensitive to different aspects of workload and are, thus, better able to answer different workload-related questions:

Sustained vs. Momentary Workload

If the goal of a workload analysis is to estimate average or total workload associated with performing a particular task, then any form of rating scale might be appropriate, as would overall performance on the most salient measures, or average heart rate. If a more fine-grained analysis of momentary variations in workload are required, then subjective ratings would be inappropriate, as would global performance measures. However, momentary variations in heart rate or performance on a continuous control task, secondary tasks (and associated ERPs) timed to coincide with specific periods of interest, and the pattern of eye movements and eye blinks might be appropriate.

Reserve Capacity

If the research question seeks to identify operators' potential abilities to perform a task for a longer duration or under more extreme environmental conditions or to assume additional responsibilities, then measures of primary task performance will provide little information. Prospective subjective ratings might be used to index the likely workload under such hypothetical conditions. Secondary tasks would provide the most sensitive measure; the level of performance operators are able to achieve on one or more secondary tasks (by maintaining a constant level of effort and performance on the primary task and by employing all of their available resources) can index the resources required by the primary task, and thus by inference, the amounts and types of resources still available to perform additional activities. ERPs associated with the presentation of secondary task stimuli would provide additional useful information.

Emotional and Physical Consequences

Heart rate and multidimensional subjective ratings (such as SWAT and NASA-TLX) provide the most sensitive indices of the stress often

associated with inappropriate levels of workload. In addition, heart rate and NASA-TLX also provide an indication of the physical demands placed on an operator. Primary and secondary task performance and ERPs would provide little information.

Specific Source of a Workload Problem

Multidimensional rating scales, such as SWAT and NASA-TLX, and a battery of secondary tasks are the best approaches for diagnosing the specific cause of a workload problem. Depending on the method chosen, the user might determine that excessive time pressure, heavy visual or manual demands, excessive mental effort, or emotional stress was the primary cause of an overall increase in workload or a performance decrement. In addition, the pattern of primary task performance decrements might also provide diagnostic information.

Comparison Among Design Alternatives

Any subjective rating scale can provide information about the workload associated with alternative designs, however multidimensional ratings scales will provide more diagnostic information. Secondary tasks designed so as to require similar resources to those also required in using the design alternative would prove useful information as well. Primary task measures of performance with the alternative systems would be essential. Patterns of eye movements and visual fixation durations might be particularly useful in evaluating alternative visual displays. ERPs and response latency might be used to index how quickly operators were able to notice or extract relevant information.

Research Environment

Although studies performed in the actual system, or an adequate simulation of it, provide the most accurate assessments of workload, it is possible to obtain useful information through research conducted in similar systems, part-task simulators, or the laboratory using more abstract representations of the activities that will be performed when it is not feasible to conduct the study using the actual system. In fact, it is often possible to obtain more precise information by evaluating the impact of specific elements of a task in isolation. However, it will be difficult to assess the influence of environmental factors in any other environment than the real thing. Different measures will be more or less appropriate depending on the constraints of the environment in which an assessment is performed.

Laboratory

In the laboratory environment, it is feasible to implement any of the measures described above. Here, the decision of which measure to use might be based on the equipment and funds available, the techniques with which the experimenter has expertise, the design of the experimental tasks, and the specific research question being addressed. Generally, both primary and secondary task measures can be obtained with great precision and ratings can be timed so as not to interfere with task performance. Although ERPs have been used successfully in this environment, heart rate is generally insensitive.

Simulation

In simulation research, large quantities of primary task performance data are generally available. Here, the problem is one of selecting which measures are relevant and summarizing and integrating the information provided by multiple measures. Subjective ratings are generally appropriate, as are most physiological techniques. Secondary tasks may be somewhat more difficult to implement, depending on the flexibility of the simulation hardware and software and the creativity of the experimenter. As the complexity of the activities being simulated increases, correlating the presentation of secondary tasks, ERPs or patterns of eye movements to the operator's activities can become quite difficult.

Field Tests

In the operational environment, accurate measures of performance are often not available, and it may be difficult to implement secondary task measures. However, attempting to obtain such measures would be worth the effort. Particularly in field tests, presenting secondary tasks as embedded elements of the overall task is particularly important. In this environment, subjective ratings are the most commonly used measure, although the experimenter must exercise care that the rating procedure does not interfere with task performance, thereby compromising safety. Heart rate measures are generally useful in an operational situation and can be obtained unobtrusively and with little difficulty using portable recording devices. ERPs, eye movements, eye blinks, and pupil diameter measures are generally not appropriate, due to practical constraints.

Estimating Imposed Demands

A preliminary analysis of objective task requirements, the equipment that will be used, and the environment in which the task will be performed is necessary to select the most appropriate measures. This analysis, which

might be quite informal, defines: (1) the magnitude of task demands, and (2) the most salient workload dimensions. In addition, this analysis can aid in identifying the performance levels operators will be expected to achieve and practical opportunities for introducing workload measures.

Magnitude of Task Demands

Since different measures are particularly effective when overall workload is generally low, moderate, or high, a preliminary analysis will increase the probability that measures will be selected and applied effectively. *A priori* assessments of imposed task demands generally catalog the number of activities that an operator will be expected to accomplish and the precision with which they must be performed. Task demands can be approximated by computing the information that must be processed, the number of decisions that must be made, the number and frequency of responses, the difficulty of mental calculations and transformations, and the predictability of events. The difficulty of almost any task can be altered by reducing the time available for performance; workload can be approximated by comparing the time required to complete subtask components to the time available. Task scheduling also influences workload; subtasks performed in sequence may impose relatively low workload, whereas unacceptably high workload may occur if they must be performed concurrently. The workload associated with the scheduling of task components can be approximated by evaluating the complexity of schedules operators must remember or develop, the frequency with which multiple tasks must be performed simultaneously, and the degree to which concurrent tasks require similar or different resources.

In general, secondary tasks are most appropriate when primary task demands are low to medium. Primary task measures are most appropriate when task demands are moderate. Subjective ratings can be used across the range of task difficulties. Heart rate is more effective for relatively demanding tasks, particularly when stress or responsibility are involved. Eye movement data and ERPs may be difficult to interpret for very complex tasks.

Sources of Task Demands

The system resources available (e.g., controls, displays, automatic subsystems, other crewmembers, support personnel) define (or limit) the task-performance strategies available to an operator. The workload impact of these factors can be approximated by assessing the types of information that will be available, the number of different information sources that must be monitored and integrated, the location and handling characteristics of controls, and the availability of automatic subsystems.

Analyzing the human resources that will be required to perform a task is particularly important in designing and implementing secondary tasks. To be most effective, secondary tasks must require the same resources that are required to perform the primary task. In addition, a thorough analysis of the primary task might suggest opportunities for introducing embedded secondary tasks. In highly automated systems, primary task measures will provide little information about workload. However, primary task measures may prove an excellent source of information when the operator is required to make frequent, measurable inputs. Multidimensional rating scales are most appropriate when assessing tasks with many, overlapping sources of workload; subscale ratings can identify variations in the composition of workload over time.

Implementing Workload Measures

The quality of information provided by a workload measure depends entirely on the quality of the study in which it was obtained. Although experimental control becomes more difficult as the research environment moves from the laboratory to simulation and field test, the additional effort required to conduct a "clean" experiment is always worth its cost.

Training

Some techniques can be administered adequately by relatively inexperienced technicians (e.g., rating scales) while others require substantial sophistication and experience (e.g., ERPs). Thus, both experimenters and subjects must be familiarized with the assessment procedures before experimental data are collected. For example, the terms used in rating scales must be defined clearly so that raters will adopt the definitions and criteria that the experimenter intended. Operators must receive advance training with secondary tasks to ensure they have reached asymptotic performance; if they have not it will be difficult to determine whether variations in secondary task performance reflect primary task workload or increased expertise with the secondary task itself. Because operators' strategies and skills change as they gain experience with a task (thereby changing their workload), it is essential that the subjects in a workload experiment are similar (in terms of experience and skills) to the people who will be the operational users of the system.

Experimental Control

Counterbalancing the order in which different conditions, tasks, displays, etc., are presented will distribute the effects of training (on the primary task)

and experience (with the evaluation procedure). If, for example, the most difficult task is always presented last, the effects of increased familiarity or improved skills (a potential source of workload-reduction) may result in erroneous conclusions. In addition, if identifying the workload associated with a specific aspect of a complex system is the goal of a workload analysis, it is essential that all other sources of workload variability are held constant across measurement intervals.

Comparing Measures Across Tasks

A major concern is that it is difficult to compare the workload of two tasks directly if different measures were obtained. The most obvious problem is that units of measurement and scaling might be different. This problem might be solved by standardizing the scores or expressing them in terms of percent change. However, since different measures are sensitive to different aspects of workload, it may be inappropriate to compare them directly. In addition, individual differences in the use of rating scales, resting heart rates, and level of experience on a task may create large differences in absolute levels between experiments that do not, in fact, represent true workload differences. And, apparently similar measures of performance may in fact have different meanings, reflect differences in systems (rather than the behavior of the human operator), or task-specific instructions and priorities across different tasks. Thus, workload measures are most useful in assessing relative differences in workload within a specific context, and less useful for comparisons across tasks and experiments.

Quantification and Analysis

In summarizing workload analyses and making recommendations, measurement variability must be reported as well as average differences. Formal statistical analyses take between-subject variability into account when determining the statistical significance of average differences. However, when data are available for only a few operators, a situation common in many operational situations, statistical treatment of the data is nearly impossible. In this case, it is imperative that the ranges of values are reported as well as the averages. Finally, just because different measures can be expressed with great numerical precision, this does not imply that the information they provide is equally precise. For example, secondary task reaction times might be recorded to the nearest millisecond, while there may be only three possible values for a rating, however, the sensitivity of both measures might be such that neither can do more than distinguish between high and low workload. However, precision of measurement, rather than precision of expression, is important; inaccurate or "noisy" data certainly limits the faith that can be placed in the information provided.

Interpretation

To be useful, workload measures must be translated into terms that are meaningful to the end users. Reporting that ratings of "25" and "35" were obtained for System A and B is not particularly meaningful. However, reporting that ratings for "System A were 25 percent higher than for the reference system" puts this information in context and compares it to a known quantity. Again, note that the example is couched in relative terms; it is still very difficult to make absolute statements about workload (e.g., "This task requires 75 percent of the operator's capacity"; "Momentary workload was excessive 25 percent of the time"). Even "objective" measures that are reported in physical units, as opposed to obviously subjective values, do not, in fact, convey information that is as concrete and absolute as a similar statement about fuel consumed or exceeded oil pressure limits. The ranges of acceptable and unacceptable workload simply are not known. Identifying these points or ranges has proven to be difficult because sources of workload vary among tasks, different people respond to the same objective demands by adopting different strategies and exerting different effort, and individuals' abilities to cope with excessively low or high workload differ.

WHY DO MEASURES DISSOCIATE?

Often, workload practitioners and system designers are confronted by a lack of agreement or instances of dissociation between different measures of workload, or between measures of workload and performance (e.g, Vidulich, 1988; Vidulich & Wickens, 1986; Wickens & Yeh, 1988): (1) The correlation between workload measures is either not significant or is negative. For example, people (or systems) that appear to have higher workload by one measure are shown to have lower workload by another; or (2) A system shown to have significantly higher workload by one measure, may have lower workload (or better performance) by another measure. An understanding of the theory of workload, as outlined earlier in the chapter, as well as an understanding of the nature of the measures used, may make many of such dissociations interpretable.

Figure 9-2 presented a number of examples which are interpreted within the framework of the performance resource function. Consider both Figures 9-2d and 9-2e, for example. In Figure 9-2d, circumstances that produced better performance – as a consequence of allocating greater effort – also led, naturally, to higher workload. In Figure 9-2e, the circumstances that resulted in higher workload (the use of the less "natural" keyboard interface) were different than those that resulted in better performance (the faster response of the keyboard system).

Given that both workload and performance are influenced by many different factors, and given also that some measures are affected by factors

that are not directly related to workload, it is not surprising that measures may dissociate. For example, the performance of certain secondary tasks might be disrupted by structural interference with the primary task (e.g., the eyes cannot look in two directions at the same time), while heart rate measures might reflect physical exertion. If the amount of physical load and the effect of structural interference are not equivalent in two systems that are compared, it is not surprising that these measures would not coincide.

To this form of dissociation may be added a note of caution about the use of correlations in comparing workload measures. Correlations may be computed between people, between systems or between people and systems, depending upon what is defined as a "case" (e.g., the data point in a scatter plot upon which a correlation is based). In the case of correlations between people, the data tell us whether the people who find a system difficult to operate (as indicated by a higher subjective rating, for example), are the same people who show low performance on the system or respond to an increase in difficulty by a measurable increase in heart rate. The major problem with the use of these correlations is that there are so many differences between people (e.g., resources available, skill on a particular task, resource-investment strategy, and willingness to use the high or low ends of a subjective rating scale). Since each of these factors can influence different measures in different directions, low correlations are not surprising (and high correlations are not always interpretable). The correlations between systems (averaged across people) are more informative; they demonstrate whether systems that appear to have low workload (based on one measure) are the same systems that appear to have low workload (based on another measure). Here, correlations tend to be more meaningful as long as the unique characteristics of the different measures are taken into account. Correlations of the third type, in which variability between people and between systems is combined, are often uninterpretable; they mix two entirely different sources of variability to produce a single number.

HOW CAN WORKLOAD BE PREDICTED?

Up to this point, the focus of this chapter has been on workload assessment; a system is designed and its figure of merit is established. In contrast, models for workload prediction are designed to make projections about the workload imposed by a system (or a system modification) which does not yet exist. Should a helicopter be designed to support a one crew or two crew flight deck? Will the requirement to operate a communications system in degraded lighting make workload unacceptable? Will the workload reduction created by introducing an automated function be significant? To answer questions such as these, designers would like to turn to predictive models; it is far more efficient and cost effective to resolve workload problems on paper, in advance, than to wait until the first system

has been fielded and then solve a problem that arises by a change in design. Many predictive models are described in Elkind, Card, Hochberg, and Huey (1989), McMillan (1989), Phatak (1983), and Wickens (1989a, 1989b) and they will not be reviewed in detail here. However, an important feature of all of these models is that they do not explicitly predict either primary task performance or resource requirements, but leave the distinction between the two purposefully unspecified. This is because, as we have noted, workload predictions for the high-demand range focus on changes in performance, while workload predictions for the low-demand range focus on changes in reserve capacity. However, the state of the art of prediction is not yet sufficiently precise to specify within which range a prediction will fall. All that is offered is a general predicted figure of merit for a system, or a time line which shows how this figure of merit might fluctuate over time.

Typically, these models are employed for one of two purposes: to make a prediction of absolute workload (e.g., whether the workload imposed on a single pilot flying a particular mission will be excessive) or to make predictions of relative workload differences among design alternatives. The foundation of most predictive workload models is a task analysis (in which the mission is decomposed into specific activities that the operator will perform) and a time line analysis which computes workload based on the percentage of time the operator will be busy performing relevant tasks, as shown in Figure 9-3 (Parks & Boucek, 1989). Usually, some value, such as 80 percent, is chosen to define a "red line," above which workload is considered to be unacceptably high. Task performance times are estimated by reference to standardized tasks in a data base, by empirical research, or expert opinion.

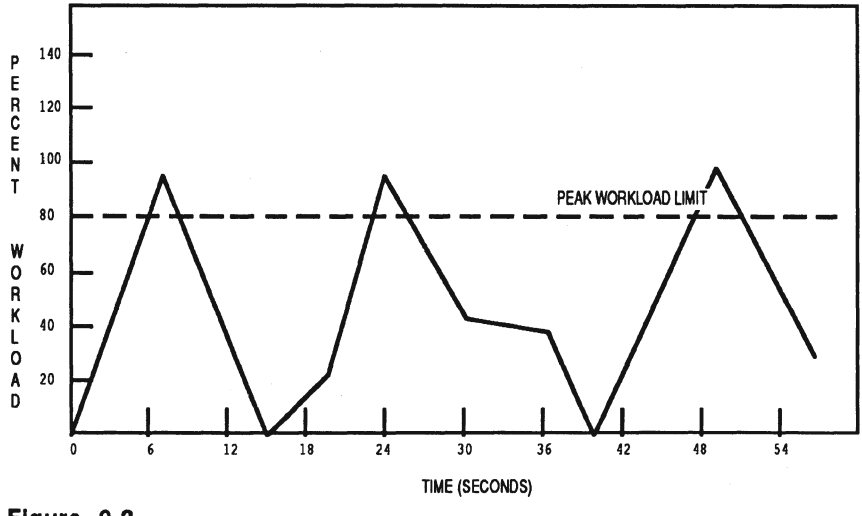

Figure 9-3
Example of Workload Time History Profile as Produced by TLAP

Two advances to time line analysis, which incorporate information about task difficulty and task interference, have added greater precision to the predictions. McCracken and Aldrich (1984) and Aldrich, Szabo, & Bierbaum, (1989), have argued for the importance of coding the resource demands of different activities in the time line, such that the workload of two easy tasks carried out concurrently might be less than (or at least not double) the workload of a single difficult task. Aldrich et al. developed a table lookup coding scheme for defining the demand values of different activities on four 7-point scales (visual, auditory, cognitive, physical load). Recent validation of predictive workload model assumptions suggests that this feature is an extremely important component of useful models (Wickens, Larish, & Contorer, 1989). For example, time-sharing vehicle control and vocal communications is fairly easy for well-trained operators, but time-sharing reading procedural instructions with communications is not. North and Riley (1989) and Wickens, Harwood, Segal, Tkalcevic, and Sherman (1988) developed predictive workload models that attempt to consider competition for multiple processing resources, as well as demand levels and time lines.

A final feature of some predictive models that is of considerable value in forecasting real world operating conditions is the ability of these models to dynamically schedule task performance according to momentary demands. A static time line, such as that shown in Figure 9-3 in which task analysis establishes the moment at which tasks will be performed, does not take into account the fact that an operator may choose to postpone the performance of a demanding task if it occurs when workload would otherwise be excessive. Models such as the Human Operator Simulator (HOS) (Harris, Iavecchia, Ross, & Shaffer, 1987; Harris, Iavecchia, & Bittner, 1988), MicroSaint (Laughery, Drews, & Archer, 1986), and Procedure Oriented Crew Model (PROCRU) (Corker, Davis, Papazian, & Pew, 1986) have incorporated this dynamic process of activity rescheduling.

To provide accurate workload predictions, a computational model must address not only single task demand levels and performance times, but also penalties for concurrent task conflicts. By combining such information, the model can identify workload peaks for specific workload dimensions, or averages across dimensions, at a given point in time or averaged over intervals of time, and extrapolate the consequences of such overload points to decrements in performance. Existing models vary in the procedures they offer for task decomposition and analysis, taxonomies of task-related and behavior-related functions, grain of analysis, availability of standardized data bases of subjective or objective values for task performance times and load levels, output format, and the degree to which workload values are related to the degree or type of subsequent performance decrements. Furthermore, the scientific basis for and empirical evaluation of these models differ substantially.

In principle, it should be possible to develop predictive models that are sufficiently accurate that empirical assessments are not also required. In

practice, however, predictive models neither can, nor should, be used alone without empirical assessments to verify their accuracy, given the capabilities of existing techniques. The predictions provided by these models should not be treated as established fact, but rather as suggested guidance. In particular, the value of predictive models is that they can make rough predictions of figure of merit which, if not 100 percent reliable, are at least considerably better than chance (or intuition). Furthermore, these models can also often objectively identify points of potentially high workload and diagnose their cause. These points can then serve as the focus for in-depth empirical evaluation, using the workload assessment techniques that were described in the section entitled "How is Workload Assessed?"

Workload predictions are not an adequate substitute for an empirical evaluation; it is the interaction between a particular operator or team of operators and the demands imposed by the task, equipment, and environment that determines the workload experienced and system performance achieved, not task demands alone. The effort that operators can (or will) exert and the task-performance strategies that they adopt determine the workload they actually experience and the level of system performance they can achieve. It is this interaction that is the focus of workload assessment.

CONCLUSION

Workload is an important, integrative concept that determines the abilities of human operators of complex systems to accomplish mission requirements, given the equipment and training that are provided and the organizational and environmental constraints that are placed on them. Workload can be measured with considerable of accuracy, however, workload prediction is a much more difficult and less precise process. A variety of subjective rating scales, measures of primary and secondary task performance, and physiological indicators have been developed, tested, and used to aid designers, manufacturers, and users in quantifying the effects of task requirements on the operators. Because each measure is especially sensitive to different workload causes and consequences, the results obtained with different measures may not covary. However, recent research in the field is focused on clarifying the underlying causes of such dissociations and formulating a model of workload/performance trade-offs to aid in interpreting the results of workload analyses, identifying workload criteria, and improving the accuracy of workload prediction.

REFERENCES

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). The development and application of models to predict operator workload during system

design. In G. McMillan (Ed.), *Applications of models to system design* (pp. 65-80). New York: Plenum Press.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276-292.

Bittner, A. C., Byers, J. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1989). Generic workload ratings of a mobile air defense system (LOS-F-H) (Technical Memorandum 1). To appear in *Proceedings of the Human Factors Society 33rd Annual Meeting*. Willow Grove, PA: Analytics, Inc.

Bortolussi, M. R., Hart, S. G., & Shively, R. J. (1989, February). Measuring moment-to-moment pilot workload using synchronous presentations of secondary tasks in a motion-base trainer. *Aviation, Space, and Environmental Medicine*, 124-129.

Byers, J. C. (1989). *Workload assessment of the pedestal mounted stinger (PMS)* (Technical Memorandum 7). Willow Grove, PA: Analytics, Inc.

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? *Advances in Industrial Ergonomics and Safety* (Vol. 1). London: Taylor & Francis.

Byers, J. C., Bittner, A. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Casper, P. A., Shively, R. J., & Hart, S. G. (1987). Decision support for workload assessment: Introducing WC FIELDE. *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 72-76). Santa Monica, CA: Human Factors Society.

Cooper, G. E., & Harper, R. P. (1969). *The use of pilot ratings in the evaluation of aircraft handling qualities* (NASA TN D-5153). Washington, DC: National Aeronautics and Space Administration.

Corker, K., Davis, L., Papazian, B., & Pew, R. (1986). *Development of an advanced task analysis methodology and demonstration for Army-NASA aircrew/aircraft integration* (Report No. 6124). Cambridge, MA: Bolt Beranek and Newman, Inc.

Derrick, W. (1988). Dimensions of operator workload. *Human Factors, 30*, 95-110.

Donchin, E., Kramer, A. F., & Wickens, C. D. (1986). Applications of brain event-related potentials to problems in engineering psychology. In M. G. H. Coles, E. Donchin, & S. Porges (Eds.), *Psychophysiology: Systems, processes, and applications* (pp. 702-718). New York: Guilford Press.

Elkind, J., Card, J., Hochberg, J., & Huey, B. (Eds.). (1989). *Human performance models for computer-aided engineering*. Washington, DC: National Academy Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87(3)*, 215-251.

Gopher, D., & Donchin, E. (1986). Workload - An examination of the

concept. In K. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance* (pp. 41-1 - 41-49). New York: Wiley & Sons.

Hancock, P. A., & Meshkati, N. (Eds.). (1988). *Human mental workload.* Amsterdam, The Netherlands: North Holland Press.

Hancock, P. A., Meshkati, N., & Robertson, M. M. (1985). Physiological reflections of mental workload. *Aviation, Space, and Environmental Medicine, 56(11)*, 1110-1114.

Harris, R. M., Hill, S. G., & Lysaght, R. J. (1989). OWLKNEST: An expert system to provide operator workload guidance. *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1486-1490). Santa Monica, CA: Human Factors Society.

Harris, R. M., Iavecchia, H. P., Ross, L. V., and Shaffer, S. C. (1987). Micro-computer human operator simulator (HOS-IV). *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1179-1183). Santa Monica, CA: Human Factors Society.

Harris, R. M., Iavecchia, H. P., & Bittner, A. C. (1988). Everything you always wanted to know about HOS micromodels but were afraid to ask. *Proceedings of the Human Factors 32nd Annual Meeting* (pp. 1051-1055). Santa Monica, CA: Human Factors Society.

Harris, R. L., Glover, B. J., & Spady, A. (1986). *Analytic techniques of pilot scanning and their application* (NASA TP-2525). Washington, DC: National Aeronautics and Space Administration.

Hart, S. G. (1986). Theory and measurement of human workload. In J. Zeidner (Ed.), *Human productivity enhancement: Training and human factors in systems design* (Vol. I, pp. 396-456). New York: Praeger.

Hart, S. G., & Hauser, J. R. (1987). Inflight application of three pilot workload measurement techniques. *Aviation, Space, and Environmental Medicine, 58 (5)*, 402-410.

Hart, S. G., McPherson, D., & Loomis, L. L. (1978). Time estimation as a secondary task to measure workload: Summary of research (NASA-CP-2060). *Proceedings of the 14th Annual Conference on Manual Control* (pp. 693-712). Washington, DC: National Aeronautics and Space Administration.

Hart, S. G., Shively, R. J., Vidulich, M. A., & Miller, R. C. (1986). The effects of stimulus modality and task integrality: Predicting dual-task performance and workload from single task levels (NASA-CP-2428). *Proceedings of the 21st Annual Conference on Manual Control* (pp. 5.1-5.18). Washington, DC: National Aeronautics and Space Administration.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam, The Netherlands: North Holland.

Hill, S. G., Byers, J. C., Zaklad, A. L., Bittner, A. C., and Christ, R. E. (1989).

Prospective workload ratings of LOS-F-H mobile & defense missile system (Technical Memorandum 2). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Zaklad, A. L., Bittner, A. C., Byers, J. C., & Christ, R. E. (1988). Workload assessment of a mobile air defense missile system. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Jex, H. R., & Clement, W. F. (1979). Defining and measuring perceptual-motor workload in manual control tasks. In N. Moray (Ed.), *Mental workload: Its theory and measurement* (pp. 125-179). New York: Plenum Press.

Kantowitz, B. H., Bortolussi, M. R., & Hart, S. G. (1987). Measuring pilot workload in a motion base simulator: III. Synchronous secondary task. *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 834-837). Santa Monica, CA: Human Factors Society.

Kramer, A. F., Sirevaag, E. J., & Braune, R. (1987). A psychophysiological assessment of operator workload during simulated flight missions. *Human Factors, 29 (2)*, 145-160.

Kramer, A. F., Wickens, C. D., & Donchin, E. (1983). A analysis of the processing requirements of a complex perceptual motor task. *Human Factors, 25*, 597-621.

Laughery, R., Drews, C., & Archer, R. (1986). A micro-SAINT simulation analyzing operator workload in a future attack helicopter. *Proceedings of the IEEE National Aerospace and Electronics Conference* (pp. 86-92). New York: Institute of Electrical and Electronics Engineers.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., Zaklad, A. L., Bittner, A. C., and Wherry, R. J. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies* (TR 851). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Mallery, C. J., & Maresh, J. (1987). Comparison of POSWAT ratings for aircraft and simulator workload. In R. Jensen (Ed.), *Fourth International Symposium on Aviation Psychology* (pp. 644-650). Columbus: Ohio State University.

McMillan, G. R. (Ed.). (1989). *Applications of models to system design.* New York: Plenum Press.

McCracken, J. H., & Aldrich, T. B. (1984). *Analyses of selected LHX mission functions: Implications for operator workload and system automation goals* (Technical Note ASI479-024-84). Fort Rucker, AL: Army Research Institute Aviation Research and Development Activity.

Moray, N. (Ed.). (1979). *Human mental workload: Its theory and measurement.* New York: Plenum Press.

Moray, N. (1988). Mental workload since 1979. In D. J. Osborne (Ed.), *International Reviews of Ergonomics, Vol. 2* (pp. 38-64). London: Taylor and Francis.

Morris, N. M., & Rouse, W. B. (1988). *Human operator response to*

error-likely situations in complex engineering systems (NASA CR-177484). Washington, DC: National Aeronautics and Space Administration.

Mosier, K. L., & Hart, S. G. (1985). Levels of information processing in a Fitts Law task (NASA CP-2428). *Proceedings of the 21st Annual Conference on Manual Control* (pp. 4.1-4.15). Washington, DC: National Aeronautics and Space Administration.

Norman, D., & Bobrow, D. (1975). Data limited and resource limited processing. *Cognitive Psychology, 7*, 44-60.

North, R. A. and Riley, V. A. (1989). W/INDEX: A predictive model of operator workload. In G. R. McMillan (Ed.), *Applications of models to system design* (pp. 81-99). New York: Plenum Press.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of Perception and Human Performance*, (Vol. 2, pp. 42-1 - 42-49). New York: John Wiley & Sons.

Ogden, G. D., Levine, J. M., & Eisner, E. J. (1979). Measurement of workload by secondary tasks. *Human Factors, 21(5)*, 529-548.

Parks, D. L., & Boucek, G. P., Jr. (1989). Workload prediction, diagnosis and continuing challenges. In G. R. McMillan (Ed.), *Applications of models to system design* (pp. 47-63). New York: Plenum Press.

Phatak, A. V. (1983). *Review of model-based methods for pilot performance and workload assessment* (Report for NASA Contract NAS2-11218). Mountain View, CA: Analytical Mechanics Association, Inc.

Reid, G. B. (1985). Current status of the development of the subjective workload assessment technique. *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 220-223). Santa Monica, CA: Human Factors Society.

Reid, G. B., & Colle, H. A. (1988). Critical SWAT values for predicting operator overload. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1414-1418). Santa Monica, CA: Human Factors Society.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185-213). Amsterdam, The Netherlands: North Holland.

Roscoe, S., & Kraus, E. (1971). Pilotage error and residual attention. *Navigation, 20*, 267-279.

Roscoe, A. H. (Ed.). (1987). *The practical assessment of pilot workload* (AGARD-AG-282, pp. 78-82). Neuilly-sur-Seine, France: Advisory Group for Aerospace Research and Development.

Roscoe, A. H., & Ellis, G. A. (in press). *A subjective rating scale for assessing pilot workload in flight. A decade of practical use* (Technical Report). Bedford, England: Royal Air Force Establishment.

Stassen, H. G., Johannsen, G., & Moray, N. P. (1988). *Internal representation, internal model, human performance, and mental workload* (EPL- 88-01). Paper presented at the International Federation of Automatic Control. Urbana-Champaign: University of Illinois, Department of Mechanical and Industrial Engineering.

Stern, J. A., & Skelly, J. J. (1984). The eye blink and workload considerations. *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 942-943). Santa Monica, CA: Human Factors Society.

Vicente, K. J., Thornton, D. C., and Moray, N. (1987). Spectral analysis of sinusarrythmia: A measure of mental effort. *Human Factors, 29 (2)*, 171-182.

Vidulich, M. A. (1988). The cognitive psychology of subjective mental workload. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 219-229). Amsterdam, The Netherlands: North Holland Press.

Vidulich, M. A., & Wickens, C. D. (1986). Causes of dissociation between subjective workload measures and performance. *Applied Ergonomics, 17*, 291-296.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance*, (Vol. VIII, pp. 239-257). Englewood Cliffs, NJ: Erlbaum.

Wickens, C. D. (1984). Processing resources in attention. In *Varieties of attention* (pp. 63-102). New York: Academic Press.

Wickens, C. D. (1989a). Models of multitask situations. In G. McMillan (Ed.), *Applications of models to system design* (pp. 259-273). New York: Plenum Press.

Wickens, C. D. (1989b). Resource management and time sharing. In J. Elkind, S. Card, J. Hochberg, & B. Huey (Eds.), *Human performance models for computer aided engineering* (pp. 180-202). Washington, DC: National Academy Press.

Wickens, C. D., Harwood, K., Segal, L., Tkalcevic, I., & Sherman, W. (1988). TASKILLAN: A simulation to predict the validity of multiple resource models of aviation workload. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 168-172). Santa Monica, CA: Human Factors Society.

Wickens, C. D., Hyman, F., Dellinger, J., Taylor, H., & Meador, M. (1986). The Sternberg memory search task as an index of pilot workload. *Ergonomics, 29*, 1371-1383.

Wickens, C. D., Larish, I., & Contorer, A. (1989). Predictive performance models and multiple task performance. *Proceedings of the Human Factors Society 33nd Annual Meeting* (pp. 96-100). Santa Monica, CA: Human Factors Society.

Wickens, C. D., & Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors, 30 (5)*, 599-616.

Wickens, C. D., & Weingartner, A. (1985). Process control monitoring: The effects of spatial and verbal ability and current task demand. In R. Eberts

(Ed.), *Trends in ergonomics and human factors* (pp. 25-32). Amsterdam: North Holland Publishing Company.

Wickens, C. D., & Yeh, Y. Y. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30 (1)*, 111-120.

Wierwille, W. W., Casali, J. G., Connor, S. A., & Rahimi, M. (1986). *Evaluation of the sensitivity and intrusion of mental workload estimation techniques. Advances in man-machine systems research* (Vol. 2, pp. 51-127). Greenwich: JAI Press, Inc.